Influences of Covariates and Baseline Period on the Performance of Regression-Based Prediction of Public Health Data

Michael W. Thompson, Ph.D.,[1] Howard S. Burkom, Ph.D.,[1] Yevgeniy Elbert, M.S.[2]

[1]The Johns Hopkins University Applied Physics Laboratory, Laurel, MD, [2]Walter Reed Army Institute of Research, Silver Spring, MD

Contact: Michael.Wayne.Thompson@jhuapl.edu

INTRODUCTION

The Electronic Surveillance System for the Early Notification of Community-Based Epidemics (ESSENCE) uses syndromic and nontraditional health data to provide early warning of abnormal public health conditions. The data streams monitored by ESSENCE (e.g., numbers of over-the-counter medication sales, outpatient visits, etc.) often exhibit characteristic short-term temporal behaviors, such as seasonal trends or fluctuations associated with the day of week. In the course of routine monitoring of the data streams for anomalies, ESSENCE alerting algorithms attempt to remove these background effects by adaptively modeling the recent historical data using multiple regression. Based on this modeling, the expected present value in each monitored time series is predicted and compared to the observed present value to determine whether the residual of prediction satisfies a statistical hypothesis test. When the test fails, the anomaly is flagged for follow-up investigation as a potential indicator of an abnormal health event.

PROCEDURE

The present investigation focuses on how the performance of the ESSENCE regression-based prediction algorithm is affected by the temporal behaviors included in the regression (the covariates) and the number of days of recent historical data modeled (the baseline period). In order to identify an optimal set of covariates and baseline period that will minimize the residuals of prediction, four different data streams of daily counts are analyzed. Each of these data streams is processed using various combinations of covariates and baseline periods, and the resulting prediction residuals during a one-year time period are recorded.

Data Streams

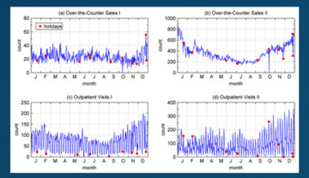

Figure 1. One-year segments of the four data streams used in the analysis are plotted in panels a–d. Blue lines indicate the number of occurrences per day (the daily count) of either over-the-counter medication sales or outpatient visits. Filled red circles indicate the count on U.S. federal holidays.

Day-of-Week Fluctuations

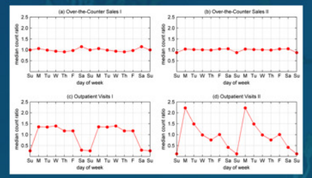

Figure 2. Panels a–d show the day-of-week fluctuations that are characteristic of the four data streams plotted in Fig. 1. The median count ratios shown here are obtained by dividing the observed daily counts by their centered 21-day moving averages, then grouping the resulting ratios according to the day of week and calculating the median value within each group (excluding holidays). Panels c and d exhibit significant day-of-week fluctuations, whereas panels a and b do not.

Table 1. Definitions of the regression covariates used in the analysis. These covariates are combined to form 10 unique covariate sets.

Symbol	Description
C	a constant
LT	a linear trend, centered on the baseline period
DOW1	a set of two indicator variables, representing three day-of-week categories
H+DOW1	a set of two indicator variables, like DOW1, but with special treatment of holidays and days after holidays
DOW2	a set of six indicator variables, representing seven day-of-week categories
H+DOW2	a set of six indicator variables, like DOW2, but with special treatment of holidays and days after holidays

Regression-Based Prediction

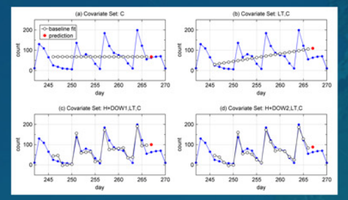

Figure 3. Each of the examples of regression-based prediction shown here is based on the Outpatient Visits II data stream (see Fig. 1d) and uses a different covariate set (see Table 1 for definitions of the individual covariates). In each of these examples, the count on day 267 is predicted using the regression covariate coefficients that provide the best fit during the preceding 21-day baseline period, which spans days 246 through 266. The non-standardized prediction residual is calculated by subtracting the predicted count from the observed count on day 267. The standardized prediction residual is calculated by dividing the non-standardized residual by the standard error of regression of the baseline period. By repeating this procedure for each successive day in the data stream, one obtains time series of the standardized and nonstandardized prediction residuals during the entire one-year period plotted in Fig. 1.

RESULTS

For each unique combination of covariate set and baseline period used, the performance of the prediction algorithm is evaluated by observing the overall scatter in the nonstandardized prediction residuals, the bias in the standardized prediction residuals when grouped by day of week, and the bias in the standardized prediction residuals on holidays.

Median Absolute Prediction Error

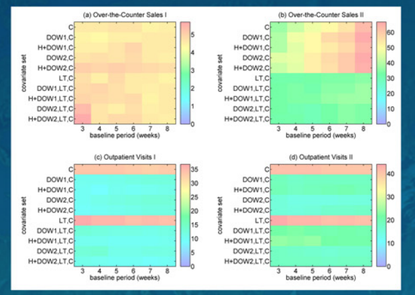

Figure 4. The error shown here is obtained by calculating the median absolute value of the nonstandardized prediction residuals during the one-year period plotted in Fig. 1. The variation of this error with respect to the covariate set and the baseline period is plotted using colored rectangles arranged in a grid.

In panel b, the use of the LT covariate reduces the error because the Over-the-Counter Sales II data stream (see Fig. 1b) contains significant seasonal trends superimposed on relatively small day-of-week fluctuations (see Fig. 2b). In panels c and d, the use of any type of DOW covariate reduces the error significantly because of the large day-of-week fluctuations that are present in these two data streams (see Figs. 2c and d).

Day of Week Having the Largest Absolute Median Prediction Error

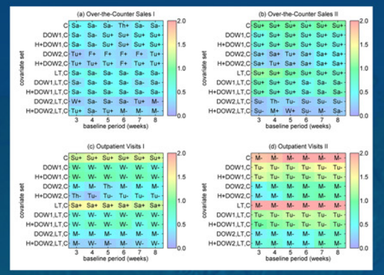

Figure 5. The error shown here is obtained by grouping the standardized prediction residuals during the one-year period plotted in Fig. 1 according to the day of week, then calculating the absolute median value within each group (excluding holidays) and selecting the day of week with the largest value. The variation of this error with respect to the covariate set and the baseline period is plotted using colored rectangles arranged in a grid. The day of week on which the maximum error occurs is notated within each rectangle, along with an indication of whether the majority of predictions on that day are greater than (+) or less than (-) the observed counts.

In all four panels, the use of either type of DOW2 covariate reduces the error, but the effect is most significant in panel d because of the complex day-of-week fluctuations that are present in this data stream (see Fig. 2d). In panels b–d, the error is also reduced when shorter baseline periods are used.

Absolute Median Prediction Error on Holidays

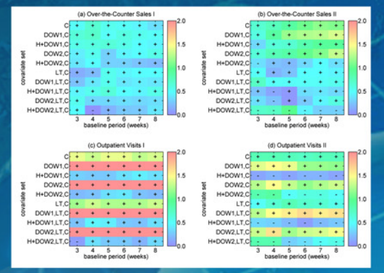

Figure 6. The error shown here is obtained by calculating the absolute median value of the standardized prediction residuals that occur on holidays during the one-year period plotted in Fig. 1. The variation of this error with respect to the covariate set and the baseline period is plotted using colored rectangles arranged in a grid. The symbol within each rectangle indicates whether the majority of predictions on holidays are greater than (+) or less than (-) the observed counts.

In panels c and d, the use of either type of holiday-based DOW covariate reduces the error, but the effect is most significant in panel c because the behavior on holidays in this data stream (see Fig. 1c) is more uniform than in any of the three other data streams (see Figs. 1a, b, and d).

CONCLUSIONS

The optimal set of covariates that minimizes the various prediction errors consists of: C, a constant; LT, a linear trend, centered on the baseline period; and H+DOW2, a set of six day-of-week indicator variables, with special treatment of holidays. The optimal baseline period is approximately four weeks. Although these results have been obtained by assuming that the regression residuals are normally distributed, the optimal parameter values are identical when the regression residuals are assumed to be Poisson distributed instead.

Funding for this project has been provided to The Johns Hopkins University Applied Physics Laboratory on behalf of the National Capital Region by the U.S. Department of Homeland Security Urban Areas Security Initiative.